自動採点システムの評価と学習者の作文に与える影響

李　在鎬（早稲田大学）
村田　裕美子（ミュンヘン大学）
スルダノヴィッチ・イレーナ（プーラ大学）

要旨

私たち研究グループでは，「jWriter」（https://jreadability.net/jwriter/）というウェブシステムを開発している。本論文では，システムの妥当性を評価する目的で行った調査結果を報告する。調査の結果，次の 2 点が明らかになった。1）日本語の運用能力を測る客観テスト「SPOT90」と「jWriter」の自動採点の結果には中程度の相関がみられ，日本語学習者の「書く力」を測るツールとして妥当であること，2）「jWriter」が出力する推敲のためのフィードバック（診断的評価）は初級レベルにおいて特に有効であることが明らかになった。

【キーワード】自動採点，書く力，人工知能，統計分析

Keywords: automatic scoring, writing proficiency, artificial intelligence, statistical analysis

1　研究背景と目的

近年，人工知能や機械学習をめぐる研究の進展が目覚ましく，人が書いた文章をコンピュータが評価する研究も盛んに行われるようになった（石井・近藤（編）2020，小森ほか 2022）。いわゆる自動採点と呼ばれる技術が日本語教育分野でも普及している。こうしたことを受け，本研究グループでは，「jWriter」というウェブシステムを開発しているが，本システムは，学習者の作文のプロフィシエンシを 5 段階（入門，初級，中級，上級，超級）に自動評価することができる（李・長谷部・村田 2019，Lee & Hasebe 2020）。本論文では，「jWriter」が行う自動採点の妥当性を確認

する目的で行った調査結果を報告する。具体的にはクロアチアのプーラ大学の日本語学習者に日本語の客観テストを受けてもらったあと，「住みやすい国の条件と理由」（村田 2021）というタイトルで意見文を書いてもらった。以下では，このデータを自動採点した結果を報告する。調査の結果，「jWriter」は日本語学習者の「書く力」を測るツールとして妥当であること，特に初級レベルにおいて有効であることが明らかになった。

2　研究課題

本研究の研究課題は，以下の 2 点である。

[研究課題 1]：「jWriter」は日本語学習者の「書く力」をどの程度，捉えられるか。

[研究課題 2]：「jWriter」のフィードバックコメント（診断的評価）はどのレベルの日本語学習者に有効に働くか。

[研究課題 1]に対しては，日本語学習者が書いた意見文を「jWriter」で自動採点し，運用能力を測定する客観テストのスコアとどの程度，相関があるかを調べた。[研究課題 2]に対しては，「jWriter」からのフィードバックを受ける前と後の文章を比較し，どのような変化があったかを調べた。

3　調査デザイン

本研究では，前節の研究課題を遂行するため，クロアチアのプーラ大学（Juraj Dobrila University of Pula）の人文学部 日本語・日本文学科の日本語学習者 26 名に対して，次の頁に示す 3 つの調査を行った。

【調査 1】：日本語の運用能力を測定する客観テスト「SPOT90」を受験。

【調査 2】：「住みやすい国の条件と理由」というタイトルで意見文を執筆。

【調査 3】:【調査 2】で書いてもらった意見文を学習者自身が自動採点システム「jWriter」に入力し，システムからのフィードバックコメントをもとに作文を再執筆。

日本語学習者の「書く力」に対する自動採点の前提として，まずは，客観テストとの関連を調べる必要があると考え，【調査 1】では小林（2015）が開発したコンピュータテスト「SPOT90」（https://ttbj.cegloc.tsukuba.ac.jp/doc/teacher/doc_jp.pdf）を受験してもらうことにした。次に，【調査 2】として，「住みやすい国の条件と理由」というタイトルで意見文を書いてもらった。なお，すべての調査データは，「住みやすいコーパス」（https://sumiyasui.jpn.org/）（村田 2021）としてウェブ上で公開している。実際に収集した作文例を以下に示す。

我々の現在の激動の時代に、世界中からの困難と争いに直面しながら、世界はどのようにより良い場所にするかをしばしば自問するようになる。教育制度、職場、経済、公共サービス、コミュニティ、環境など、どんあ部門を見ても、いくつかの考えが浮かぶ。本稿では、このトピックに関して自分の考えを簡単に紹介してみたいと思う。

まず、すでに存在している手当、給付や施設を見てみると、どのように改善でき、どのように他の分野に適用できるかを考えると、住みやすい国の鍵になる。公共インフラ、教育、国民皆保険、様々なサービスとヘルプ・プログラム、および文化プログラムは全て、快適な生活を支える。また、私にとって、もう一つの最も重要なものは、コミュニティの一員としての感覚。隣人を信じられ、頼ることができ、強い相互の絆をもち、これは最終的に全て絡み合い、強力なコミュニティを作成するものである。しかも、現在に重宝なものだけでなく、将来の世代にも有益になる可能性を秘めている。また、恵まれない人々、即ちマイノリティ、貧困者、障害者などのことを考えなければならない。彼らの人生はすべて、独自の経験と知識を表しているものである。これらの過小評価されたグループを社会に含めれば、新しい創造的なアイデアと解決策が生まれるだろう。

とは言うものの、これにおいての最善の考え方は、自分が影響を与えられることから始めるということであるかもしれない。換言すれば、我々と我々の自身の行動である。最後に、地球規模の問題に対し、無力感を感じるかもしれないが、少しずつ力を合わせ、必ず住みやすい国を作ることができる。

（原文ママ）

最後に，【調査 3】として，学習者自らが「jWriter」に自身が書いた文章を入力し，システムからのフィードバックをもとに，書き直した作文を送信してもらった。「jWriter」が出力するフィードバックの例を図 1 に示す。

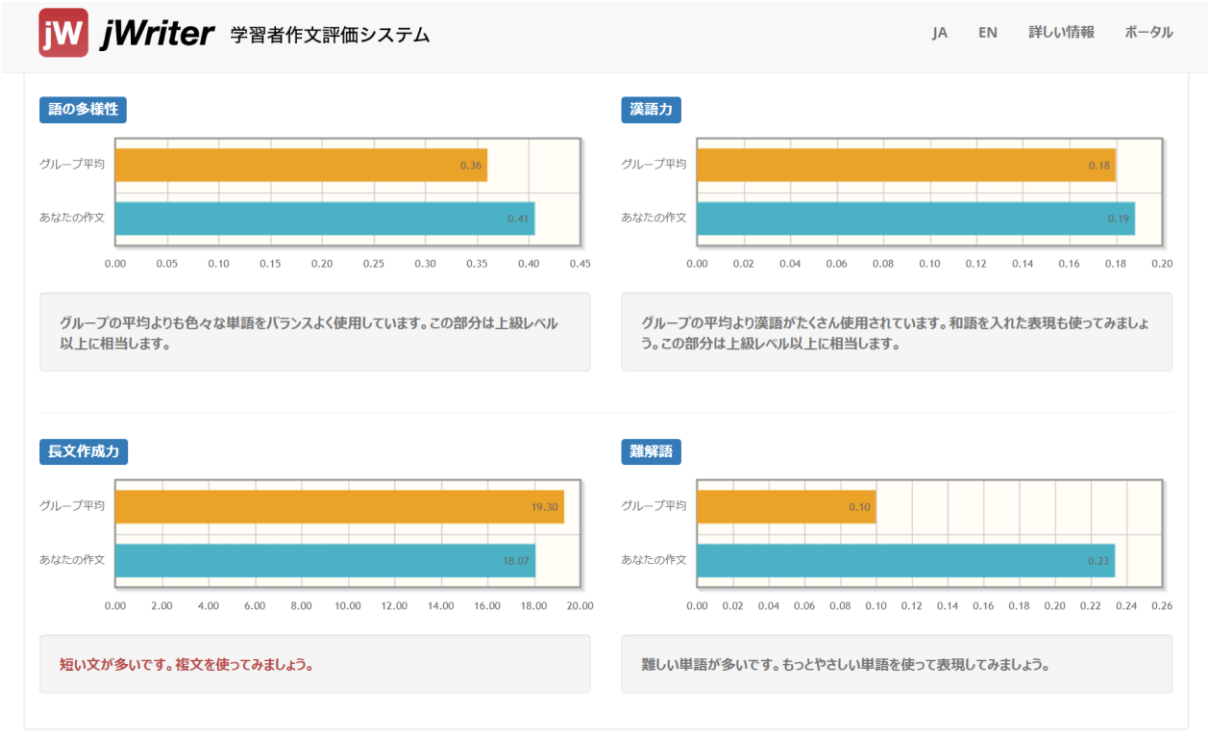

図 1：「jWriter」によるフィードバックの例

日本語学習者が「jWriter」のテキストボックスに自身の文章を入力し，「実行」ボタンをクリックすれば，図 1 のようなフィードバック（診断的評価）が受けられる。フィードバックの仕組みについては，Lee & Hasebe (2020) を参照してほしい。なお，【調査 2】と【調査 3】の学習者からのデータ送信はすべて Google フォームを用いて行った。

以上の手続きで収集したデータを「IBM SPSS（ver26)」を使用して定量的に分析した。

4　結果

【調査 1】の「SPOT90」のスコアを小林（2015）の「得点の解釈」に従って評価した結果，調査協力者 26 名は，初級が 5 名，中級が 19 名，上級が 2 名となった。この結果を踏まえ，[研究課題 1]について調べるため，【調査 2】と【調査 3】の意見文を自動採点した。図 2 に，その散布図を示す。

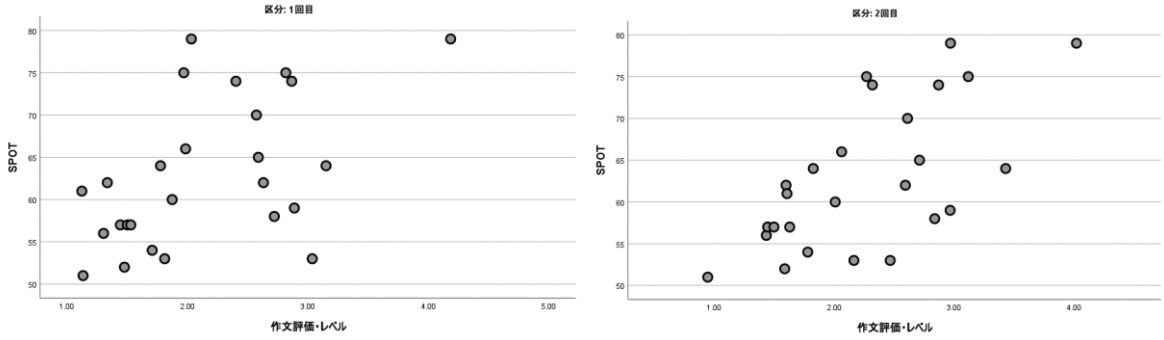

図 2：「SPOT90」のスコア×「jWriter」の自動採点スコアの散布図

図 2 の縦軸は，「SPOT90」のスコア（0 点～90 点）であり，横軸は学習者の意見文を「jWriter」で自動採点したスコア（0.5 点～4.5 点）である。図 2 の左が【調査 2】で収集した作文（1 回目の意見文）であり，右が【調査 3】で収集した作文（2 回目の意見文）である。どちらの図からも「SPOT90」のスコアが上がるにつれ，自動採点のスコアも上がっていく様子が確認できる。このことをさらに確かめるため，Pearson の相関係数を計算した。計算の結果，【調査 2】の意見文（1 回目）では「r=.518」，【調査 3】の意見文（2 回目）では「r=.653」となり，システムによるフィードバックに基づいて書き直した意見文のほうが客観テストとの相関が強くなっていることが確認できた。

以上の結果から，[研究課題 1]に対しては，次のことが明らかになった。日本語の運用能力を客観的に評価する「SPOT90」のスコアと，自動採点で意見文を評価する「jWriter」のスコア間には強い相関があり，「jWriter」は学習者の言語運用能力を適

切に捉えていると考えられる。なお，関連する研究として，李・村田・長谷部（2023）では，日本語学習者100名の意見文を自動採点し，「SPOT90」のスコアとの相関係数が「*r*=.688」であることを報告しており，本研究の【調査 3】と概ね同じ値である。

次に，[研究課題 2]について調べるため，【調査 2】で収集した意見文と【調査 3】で収集した意見文を量的分析の方法で比較した。

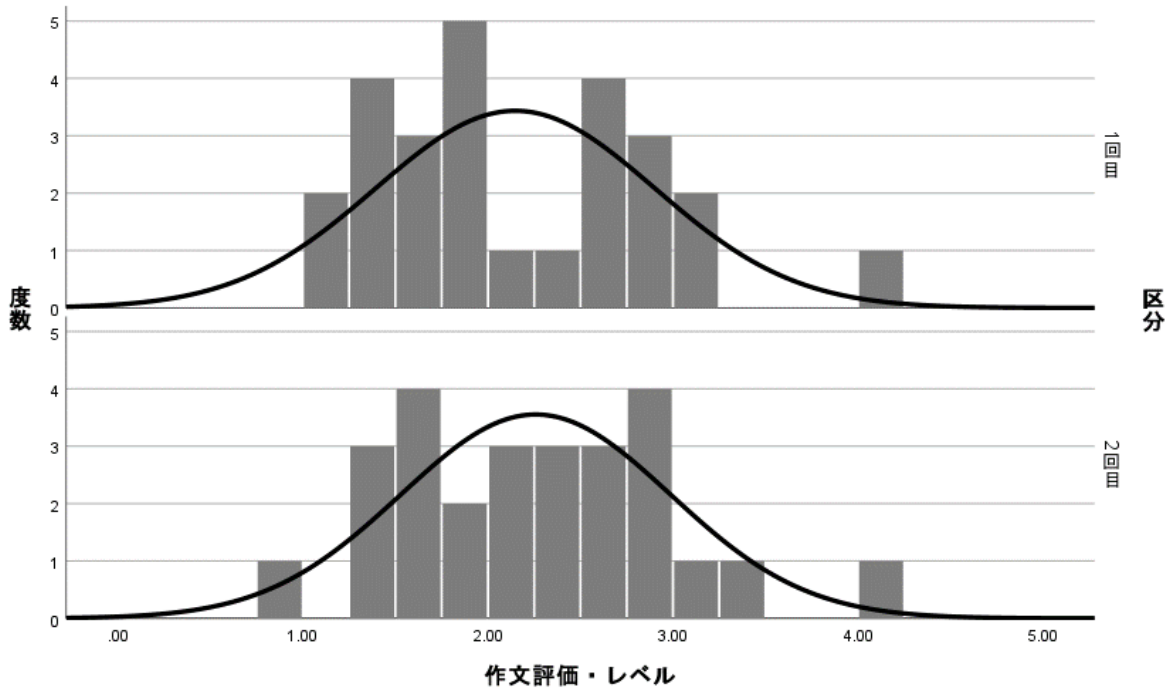

図 3：自動採点スコアのヒストグラム

図 3 の上は【調査 2】（1 回目）の意見文の自動採点スコアであり，下は【調査 3】（2 回目）の意見文の自動採点スコアである。横軸の値は 0.5 点～4.5 点の範囲で分布しており，値が高ければ高いほどプロフィシエンシが高いことを意味する。縦軸は人数を表す。図 3 を解釈すると，システムからのフィードバックを受ける前である【調査 2】（1 回目）の意見文では，1 点から 2 点の範囲で分布する低得点の集団と，3 点から 4 点の範囲に分布する高得点の集団に分かれている。一方，システムからのフィードバックを受けて書き直した【調査3】（2回目）の意見文では，低得点の集団

の一部が，高得点の集団にスライドしていることが確認できる。このことをさらに詳しく分析するため，箱ひげ図で示してみた（図 4）。

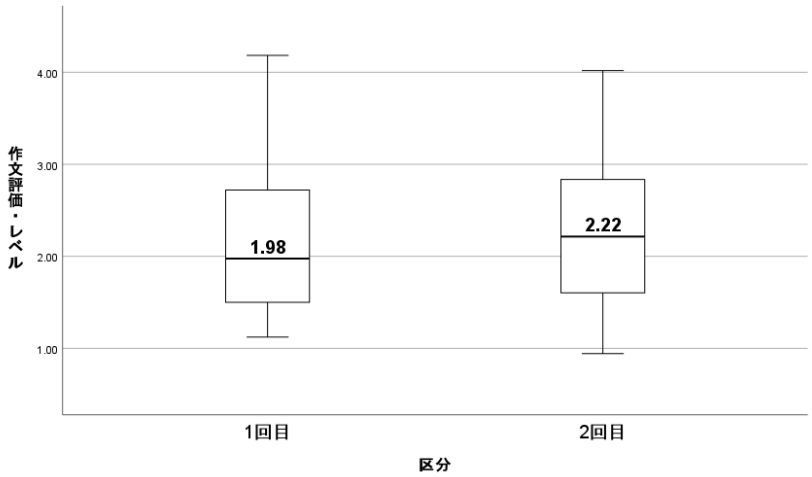

図 4：【調査 2】（1 回目）と【調査 3】（2 回目）の意見文の箱ひげ図

図 4 の中央値をみると，【調査 3】（2 回目の意見文）のほうが，わずかにスコアが上がっていることが確認できる。なお，ウィルコクソンの符号付き順位検定で統計的な有意差を確認したところ，有意であった（V=72.0, p=0.015）。次に，フィードバックによって作文のレベルにどのような変化があったかを調べてみた。

表 1：調査×作文レベルのクロス表

		作文レベル				合計
		1初級	2中級	3上級	4超級	
区分	調査2(1回目)	6	10	9	1	26
	調査3(2回目)	4	12	9	1	26
合計		10	22	18	2	52

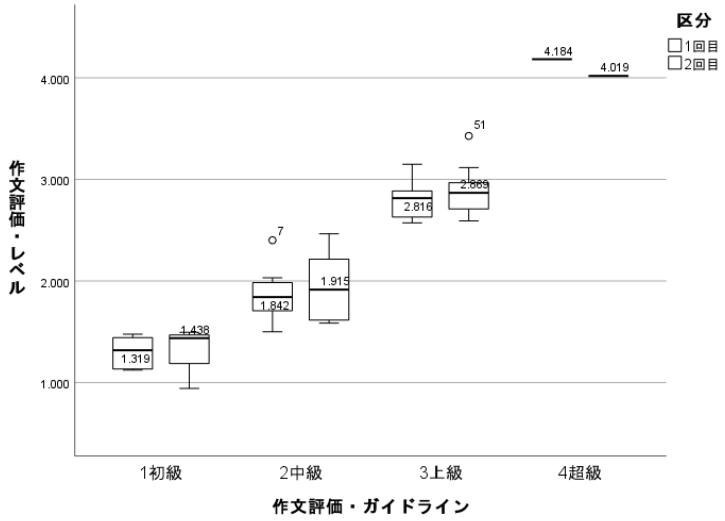

図 5：意見文のレベル別の自動採点スコアの箱ひげ図

表 1 および図 5 から上級と超級では変化がなく，初級と中級においては，レベルの変化が確認できる。このことから，[研究課題 2]に対する答えとして，「jWriter」のフィードバックコメント（診断的評価）は，初級学習者のプロフィシエンシ向上に有効であることが明らかになった。

5　まとめと今後の課題

本研究の研究課題とそれに対する答えを示す。

［研究課題 1］：「jWriter」は日本語学習者の「書く力」をどの程度，捉えられるか。

→答え：運用能力を測る客観テストのスコアとも相関が高く，日本語学習者の「書く力」を測定するツールとして概ね妥当である。

［研究課題 2］：「jWriter」のフィードバックコメント（診断的評価）はどのレベルの日本語学習者に有効に働くか。

→答え：初級学習者のプロフィシエンシ向上において有効に作用する可能性がある。

本研究の結果には，次の 2 点の問題があり，結果の有効性に関しても限定的である。

1）統制群なしの調査であること

2）サンプルサイズが小さいこと

1)の問題があるため，[研究課題 2]の結果が「jWriter」の効果であるかどうかについては確定的なことが言えない。2)の問題があるため，[研究課題 2]の結果が誤差の範囲である可能性が排除できない。これらの課題をクリアするため，今後は，ChatGPT（https://chat.openai.com/）など他のシステムを使い，より多様なフィードバックを生成させ，「jWriter」のものと比較すると同時に，調査協力者を増やし，誤差の可能性を排除することに注力したい。最後に，今回の分析はシステムの妥当性を検証することが主目的であったため，定量的分析を行ったが，学習者が書いた文章を目視で観察し，どのような部分で変化があったのかを調べる分析も必要であると認識しており，今後の課題としたい。

＜謝辞＞

本調査に協力してくれたプーラ大学の日本語学習者 26 名に感謝する。本研究は，科研費（19K21637，19H01273）の成果である。

＜言語資源・ソフトウェア＞

「jWriter」https://jreadability.net/jwriter/　（2023 年 9 月 28 日）

「住みやすい国コーパス」https://sumiyasui.jpn.org/　（2023 年 9 月 28 日）

「SPOT90」https://ttbj.cegloc.tsukuba.ac.jp/doc/teacher/doc_jp.pdf　（2023 年 9 月 28 日）

＜引用文献＞

石井雄隆・近藤悠介（編）（2020）『英語教育における自動採点 現状と課題』ひつじ書房.

小林典子（2015）「SPOT」, 李在鎬（編）『日本語教育のための言語テストガイドブック』pp.110-126, くろしお出版.

小森和子・伊集院郁子・李在鎬（2022）「日本語学習者の作文における自動評価と教師評価の比較」『明治大学国際日本学研究』14(1), pp.41-68, 明治大学国際日本学研究科.

村田裕美子（2021）「小規模コーパスの構築方法」, 李在鎬（編）『データ科学×日本語教育』pp.34-53, ひつじ書房.

Lee, J., and Hasebe, Y. 2020. Quantitative Analysis of JFL Learners' Writing Abilities and the Development of a Computational System to Estimate Writing Proficiency. *Learner Corpus Studies in Asia and the World, 5*. 105-120.

李在鎬・長谷部陽一郎・村田裕美子（2019）「学習者作文の習熟度に関する自動判定と Web システムの開発について」, 李在鎬（編）『ICT×日本語教育』pp.38-53, ひつじ書房.

李在鎬・長谷部陽一郎・村田裕美子（2023）「日本語学習者作文評価システム「jWriter」の自動採点の精度」『CASTEL/J 2023 予稿集』pp.116-119, 日本語教育支援システム研究会.